Stethoscope: Crypto Phones with Transparent & Robust Fingerprint Comparisons using Inter Text-Speech Transformations

Maliheh Shirvanian*
Visa Research
Palo Alto, CA, USA
mshirvan@visa.com

Nitesh Saxena
University of Alabama at Birmingham
Birmingham, AL, USA
saxena@uab.edu

***Abstract*—Crypto Phones are emerging apps aimed for end-to-end secure communications. To detect man-in-the-middle (MITM) attacks, traditional Crypto Phones rely upon end-users to verbally exchange and compare a short protocol fingerprint. This requirement is often found to be inconvenient by the users. Hence, most current apps do not mandate fingerprint validation, allowing the users to opt-out, completely disregarding security in favor of usability. Besides, speaking the fingerprints is not free of user errors, which may lead to rejection of benign sessions degrading the user experience.**

In this paper, we address these fundamental problems by introducing *Stethoscope*[1], a new Crypto Phone model that removes the human user from the loop of fingerprint comparison by using text-to-speech and speech-to-text transformations. Stethoscope automatically performs two tasks on behalf of the user: (1) creating the fingerprint by incorporating a fingerprint speaking tool at the sender side, built on top of a *limited-domain text-to-speech* engine, and (2) decoding/comparing the fingerprint at the receiver side based on a robust *speech-to-text* engine. Like the traditional design, Stethoscope relies on the receiver to manually verify the sender's voice to detect sophisticated voice attacks.

On the sender side, we design an automated fingerprint speaking tool based on a limited-domain text-to-speech system using reordering of words in a *phonetically-distinct word dictionary* previously spoken by the user. This tool asks the users to speak all the words in the fingerprint dictionary only once to train the system. On the receiver side, to decode the fingerprint, we design a robust speech-to-text transcription method. We evaluate the effect of automating the fingerprint creation, transfer, and comparison in the Stethoscope design against manual speaker verification with a user study. Our results show that Stethoscope provides a *0% false accept and 0% false reject rate* for the fingerprint comparison, while offering a higher level of speaker verification performance compared to traditional Crypto Phones.

***Keywords*—VoIP security, end-to-end encryption, SAS validation, key exchange validation, mobile app security**

I. Introduction

Voice (text and video) over IP is one of the most popular communication methods deployed today. To secure these types of communications against man-in-the-middle (MITM) attacks [4], [6], parties involved in the protocol should agree upon a cryptographic key to secure the communications. This key agreement should preferably not rely on a centralized key management system, as it might get compromised or be coerced by higher authorities [1]–[3], [36]. End-to-end encrypted voice and messaging apps, commonly referred to as "Crypto Phones", such as WhatsApp [18], Viber [17], and Silent Circle [15], aim to establish such end-to-end secure voice (and text) communications based on a human-centric key exchange validation mechanism.

Crypto Phones run a peer-to-peer key exchange protocol [7], [34], which generates a short code (e.g., 16-bit or 2-word), called a fingerprint, per party, with the intrinsic property that if a MITM attacker attempts to interfere with the protocol, the fingerprints do not match. To verify the equality of the fingerprints, Crypto Phones rely upon the end users to recite the fingerprint displayed on their respective devices to each other and compare the received fingerprint with the one computed and displayed locally, to ensure that the MITM attacker does not interfere with the protocol messages and compromise the protocol security (referred to as the data MITM attack [31]). Some of the apps may also ask the users to verify each others' voice to detect sophisticated voice-based MITM attacks (voice impersonation) as introduced in [29].

The requirement for the users to speak and to compare the fingerprint, may be found cumbersome and might prevent users from running the fingerprint validation protocol at all (such an opt-out or skip-through behavior has been demonstrated in many security contexts [19], [25]). Furthermore, recent research studies have shown that even if the users attempt to run the protocol the fingerprint comparison task would be prone to human errors making Crypto Phones vulnerable to data MITM attacks. Besides, manually speaking the fingerprints is not free of user errors, which may lead to rejection of benign sessions or may prolong the protocol. Therefore, it is essential to ease the path to enforce fingerprint comparison and elimination of errors therein, ideally by removing the human from this task.

In this work, we respond to this crucial problem in the current deployments of Crypto Phones with *Stethoscope*, a fully automated and transparent approach to Crypto Phone fingerprint comparison that uses inter speech/text transformations. First, Stethoscope employs text-to-speech synthesis to build fingerprint from a voice dataset of the phonetically distinct words spoken by the user and automatically transfers it over the established call channel to the other end. Second, Stetho-

*Work done at UAB.

[1]STeThoSCoPe represents "Speech-to-Text Text-to-Speech Crypto Phones"

978-1-7281-3265-5/19/$31.00 ©2019 IEEE

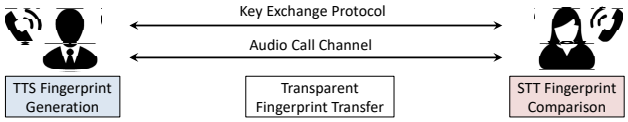

Fig. 1: Stethoscope main components include automated fingerprint generation based on text-to-speech (TTS) and automated fingerprint comparison based on speech-to-text (STT). The fingerprint is composed of words that are phonetically distinct and is transferred on the voice call channel.

scope deploys an automated fingerprint comparison based on transcription to verify the generated fingerprint.

Figure 1 shows the main components of this system at a higher level (a detailed visualization is provided in Figure 3). In Stethoscope, users speak the atomic units of the fingerprint (the words available in the fingerprint dictionary) only once. This audio dataset is then stored on the device and serves as the training dataset to a *limited domain text-to-speech tool*, which composes any fingerprint by simply putting the fingerprint words from the dataset together. Stethoscope limited domain text-to-speech tool generates the audio fingerprint per each session (in each of the user's voice) by concatenating the fingerprint pre-recorded words. Stethoscope *transparently* inserts and transfers the generated audio fingerprint over the voice channel to the other party for automated comparison. On each device, the transcriber then converts the spoken fingerprint received from the other party to text (written fingerprint), and automatically compares it with the locally computed one. The Stethoscope transcriber refers to a *phonetically-distinct word dictionary*. Stethoscope provides following key advantages over the traditional design:

1) Stethoscope enforces the fingerprint validation protocol with no user involvement, therefore, users may not opt-out or skip-through the crucial fingerprint validation task. If users opt-in to the security protocol, the security level is obviously improved compared to when they opt-out.
2) Stethoscope eliminates the potential for human errors in the fingerprint comparison task, which reduces/eliminates the chances of data MITM attacks.
3) Stethoscope eliminates the user errors in reading and/or speaking the fingerprint, which may result in rejection of a valid communication session and re-execution of the fingerprint validation protocol, and therefore leading to poor user experience.
4) Since the users are only involved in speaker verification (and not fingerprint generation and fingerprint comparison), the accuracy of the users in verifying the speaker may improve due to pure single-tasking (unlike traditional design, which are multi-tasking). Besides, in Stethoscope, longer fingerprints may be incorporated since manual speaking and comparison is not needed, which will improve the security guarantees provided by the underlying protocol.
5) The fully transparent design may improve the user experience, because the users are no more required to speak and/or compare the fingerprint and the only remaining human user task is the optional speaker verification.

Contributions: Our paper offers the following contributions:

1) **A New Crypto Phone Design for Fully Automated Fingerprint Comparison:** We introduce the Stethoscope design, which automates fingerprint comparison protocol (fingerprint speaking, exchange and comparison). Stethoscope incorporates limited domain text-to-speech technology to generate the fingerprint, transparently transfers the fingerprint over established voice call channel, and automatically compares and verifies the fingerprint using speech-to-text transcription. Stethoscope collects and stores fingerprint words spoken by the users, and uses a speech synthesis tool to build fingerprint in the user's voice, by permuting pre-recorded fingerprint word audio samples. The system starts with a fixed set of words in the dictionary but can get updated frequently to avoid possible voice attacks. Stethoscope design involves the automated transcription-based fingerprint comparison, for improved security. The Stethoscope fingerprint dictionary consists of carefully chosen *phonetically-distinct words*, the PGP word list, with different set of words for even and odd positions to achieve high transcription accuracy.
Like traditional Crypto Phones, Stethoscope relies on the receiver to manually verify sender's voice to detect sophisticated voice MITM attacks. While we do not claim to directly improve the robustness of speaker verification against the voice imitation attacks, Stethoscope helps improve the performance of manual speaker verification over the traditional designs of Crypto Phones, as supported by our experiments (introduced next). Our intuition is that this improvement stems from the use of a longer fingerprints and automating the fingerprint comparisons (thereby reducing the cognitive burden on the users).
2) **A Study to Evaluate Stethoscope Fingerprint Transcriber:** To evaluate our transcription mechanism in a realistic setting, we set up a VoIP system using FreeSWITCH [10] telephony system and transfer the audio samples over the audio call by incorporating OZEKI VoIP SIP SDK [14], with no user involvement. We evaluate our automated fingerprint comparison tool, built on top of IBM transcription technology in transcribing Stethoscope generated fingerprint samples. We collected audio samples from multiple speakers, speaking all the words in the PGP word list. We used this audio dataset to generate 40 4-word and 40 8-word fingerprints using the limited domain text-to-speech and transcribed them with the automated tool. Our transcription results show that the use of phonetically-unique PGP words *fully eliminates the errors* in the fingerprint comparison task in both benign and adversarial settings, which shows significant improvement over the general-purpose transcribers proposed in [31].
3) **Evaluation of Manual Speaker Verification:** Although this paper's goal is not to directly address the problem of voice imitation attacks against Crypto Phones , to evaluate the accuracy of the users in verifying speakers from automatically generated fingerprint samples, we ran a study and recruited 36 participants and asked them to listen to the speaker's voice to get familiar with the voice, and then listen to audio samples of the original speaker (generated by the limited domain speech to text tool) and imitated voice (synthesized by using a voice conversion tool that maps the attacker's voice to the victim's voice). The study results generally suggest that the performance of users in the speaker verification task is higher that the traditional Crypto Phones perhaps due to the reduction in user's cognitive related to the automation.

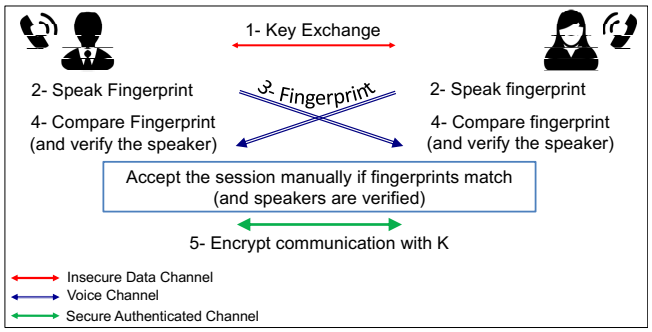

Fig. 2: Traditional Crypto Phone with manual fingerprint verification

II. Background and Related Work

A. Traditional Crypto Phones: Protocol and Threat Model

Traditional Crypto Phones are voice and messaging applications, which deploy a purely end-to-end encrypted voice and data channel. Crypto Phones adopt human-centered fingerprint validation protocols (e.g., [21], [23], [26], [27], [34]) to exchange the cryptographic key used to secure voice and data communicated over IP, followed by a human-based authentication of the key over an auxiliary/authenticated channel. Crypto Phones key exchange protocol results in a fingerprint of the protocol that is encoded into a short string of words, numbers, or sentences. To validate the fingerprints, users are supposed to exchange and compare them over an authenticated channel [33] as shown in Figure 2. In this work, we only target fingerprint exchange over voice channel as it does not require trust on any additional channel.

The devices involved in the fingerprint validation protocol are assumed to be trusted. However, the data channel over which the key is exchanged is fully controlled by an MITM attacker. The most viable attack against fingerprint validation is an MITM attack on the messages transferred on data channel as part of the key exchange protocol. As a result, the fingerprints on the two sides of the protocol do not match. Since the users speak and verify the fingerprints, they are expected to detect mismatching fingerprints (attacked channel). If the users erroneously accept mismatching fingerprints the security would be compromised, and all the voice and data communication would be vulnerable to eavesdropping or MITM.

Crypto Phones assume that the human voice channel has the property of self authentication, i.e., users can recognize each others' voice. Therefore Crypto Phones implicitly expect the users to verify the speaker's voice. However, authors of [29] introduced a voice MITM attack against Crypto Phones in which the attacker uses voice synthesis techniques (e.g., [9], [12]) to create fingerprint audio samples that sound similar to the user's voice and transfers such decorously sounding (yet bogus) fingerprint to the users. If the users can not distinguish between the synthesized voice and the legitimate user's voice the security would be jeopardized.

B. Semi-Automated Fingerprint Comparison

In a recent study [31], authors aimed to improve the security of the standard Crypto Phones by utilizing speech transcription technology. In their work, the authors designed and implemented a semi-automated fingerprint comparison tool based on standard speech-to-text engines. This model still requires the user to announce the fingerprint to the other party similar to the standard approach. However, rather than requiring the users to compare the fingerprint manually, the system automatically transcribes the spoken fingerprint and performs the comparison. Using this technique, they could reduce the possibility of accepting attack sessions from about 30% to 0%. Automating fingerprint comparisons offers other key advantages over traditional designs, including 1) use of loner fingerprint which further improves the security of the system theoretically and practically, since the fingerprint comparison is automated, 2) improving speaker verification by reducing cognitive burden on the users, thereby lowering the potential of human errors, and 3) improving the usability by automating part of the user's tasks.

Even though the study of [31] reduces the human errors in comparing the fingerprint by somewhat automating the process, it only unburdens the human user from fingerprint comparison. However, it still requires the users to manually run the protocol and speak the fingerprint over a voice call. This might prevent the users from performing the fingerprint validation task. Research studies show that the majority of the users do not validate the fingerprint, even when they are aware of the importance of this security task [32]. Furthermore, the chances of falsely rejecting legitimate sessions in [31] was still reported to be on the high side for practical usage (between 20-60%), which means that users will be forced to re-run the protocol thereby lowering the usability and adoptability. In this work, we try to address the above issues not only by automating the fingerprint comparison but also by automating the whole fingerprint validation process including the fingerprint speaking, transfer, and comparison.

III. Stethoscope Overview and Design

A. Limited Domain Text-to-Speech Translation

In our application, generating the fingerprint in the exact user's voice (not in any general TTS voice) is essential, since just as in Crypto Phones, the security is provided on the basis of human voice self-authenticity. Otherwise, an attacker who compromises the key exchange channel (data MITM) can succeed by inserting a matching fingerprint in any voice (a naive voice MITM). To generate such samples in the users voice, we could use speech synthesis, which refers to tools and mechanisms that generate machine spoken language on the basis of written input in human voice [5], [11], [13]. To train speech synthesis systems to speak in a given voice, the system should be trained with the target voice. To produce naturally sounding speech, these systems require hours of training data. However, in some application, the range of spoken output is limited to certain utterances. In such applications, limited domain text-to-speech tools can be used, which rather than recording all the utterances, record only some words and phrases and combine it with a diphone database to allow better results for common phrases and some coverage for less common phrases [20].

In our application, we essentially require the system to produce only the fingerprint samples in the user's voice. The fingerprint dictionary consists of a limited number of words (e.g., 512 PGP words). Besides, here, there is no need for

TABLE I: Comparison between Traditional Crypto Phones, Semi-Automated Design [31], and Stethoscope

	Traditinal Crypto Phones	Semi-Automated Design [31]	Stethoscope
Initiating the Protocol	Manual	Manual	**Automated (Skip-through not possible)**
Speaking the Fingerprint	Manual	Manual	**Automated (TTS)**
Fingerprint Comparison Method	Manual	Semi-Automated (general-purpose STT)	**Automated (specialized STT)**
Fingerprint Comparison Error Rate	High	Low	**None**
Speaker Verification Method	Manual	Manual	Manual
Speaker Verification Error Rate	High	Moderate	Moderate
Fingerprint Size Supported	Short	Long	Long

intonation as there is no specific context and emotions to be captured in the speech melody. Therefore, in such an application, with very limited language structure and words, we can have all utterances to be pre-recorded. The user only needs to record the words once to build a dictionary of utterances in any given fingerprint (fingerprint atomic unit dictionary). Using this audio dataset, the system can mix and match the words to map the fingerprint into an audio representation for each new session. The system picks the audio files related to each word from a pre-collected audio dataset of the user speaking the fingerprint atomic units and concatenates them to generate a whole new fingerprint.

Protecting the security of the fingerprint audio dataset is an important aspect of our system. We assume that the dataset is stored on the trusted user's device (the same way the encryption keys are stored). However, we should also make sure that an attacker cannot collect the fingerprint words spoken in the users' voice from other public channels (e.g., by attempting to establish Crypto Phones calls and saving the fingerprint words). This protection is important to defeat a Voice Reordering attack in which, an attacker collects all words in the dictionary spoken by the user and creates a legitimate sounding fingerprint. Deployment of our approach based on a "static" dictionary of words may make it vulnerable to such attack, while picking the fingerprint words from a large, "dynamic" dictionary could prevent this attack. However, since Stethoscope requires the users to speak the words in advance to build the fingerprint, having a large and dynamic dictionary would be tedious if the users are to speak all the words at once. Our mitigation is to build the fingerprint dictionary gradually. To begin, Stethoscope can ask the users to read and record a set of predefined words (e.g., a total number of 16 words to map every 4-bits to one word, which based on our experiment takes only about 20 seconds to speak), and add words to it regularly. For example, the system can ask the users to read and record a new set of words to replace those words that have been used frequently in previously generated fingerprints upon software update. To use a dynamic dictionary, Stethoscope should make sure that the users engaged in a newly established session have the same dictionary. To achieve this, the protocol can first exchange the dictionaries so that the devices agree on a subset of the fingerprint words (similar to a cryptographic protocol parameter negotiation phase, like TLS negotiation).

B. Special-Purpose Speech-to-Text Transcription

Speech to text technology takes audio content and transcribes it into written words. Speech-to-text services split the audio into utterances by silence, and then associate these smaller samples with simple phonemes. AI algorithms are then used to predict the word or phrase from the series of phonemes based on the context of the speech. Speech to text software has a near real-time efficiency and can translate thousands of words in a fraction of a second with a high accuracy [8].

In [31], authors introduced the use of speech-to-text transcriber in the context of Crypto Phones fingerprint validation. The idea behind this work is for the users to speak the fingerprint, while the devices executing the protocol can receive the audio and transcribe it to textual fingerprint. The textual presentation can then be compared with the locally computed fingerprint by the program itself without involving the user in the comparison part. This work has shown that automation of fingerprint comparison can eliminate the false accept rate and reduce the false reject rate by removing the human errors and thereby can significantly improve the security of Crypto Phones. The dictionary used in [31] is picked from a phonetically balanced voice dataset [24] containing an ordinary conversation language. Therefore, the words in their dictionary are not necessarily phonetically distinct. This choice is reflected in their reported false rejection error rate of 24.57% and 63.17% for 4-word and 8-word fingerprints, respectively. In contrast, in Stethoscope, we select words that are generally transcribed more accurately and hence, compared to the work of [31], Stethoscope transcription can offer higher accuracy. Moreover, in Stethoscope the selection of words for even and odd positions is different, allowing Stethoscope to detect transcription word deletion and insertion errors.

Recently, attacks on speech-to-text on virtual assistant apps have been introduced (e.g., [22]). These attacks use samples that are not understandable by users but are recognizable by virtual assistant apps and force the virtual assistant apps (e.g., Siri) to run malicious commands. Since Crypto Phones ask the users to verify the speaker, we expect such synthetic audio samples to be detected by users during the speaker verification. Besides, many of these attacks should have the knowledge of the phone hardware specification and can only work in the proximity of the phone [28], [35], therefore, are not applicable to Crypto Phone application.

C. System Overview and Properties

We propose Stethoscope, a new Crypto Phone model that uses text-to-speech to generate fingerprint in the user's voice based on the pre-recorded audio samples of the user speech, transfers the fingerprint over an established voice call, and automatically compares the fingerprints using speech-to-text technology. The call over which the fingerprint is transferred can be established on behalf of the user (for example upon adding a new contact). In this approach, Stethoscope app makes a transparent call to the newly added contact to establish the secure communication. The fingerprint can be played in real-time or offline (e.g., leave a voice message containing the fingerprint) for the user to manually verify the speaker. In

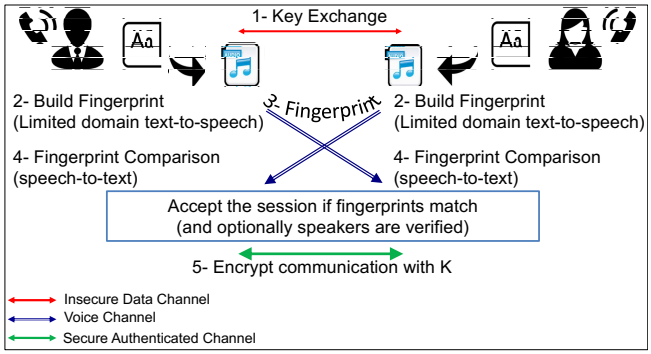

Fig. 3: Overview of the Stethoscope approach. The key exchange protocol results in a fingerprint. The audio fingerprint is automatically built from a collection of pre-recorded fingerprint words using a limited domain speech synthesizer. The fingerprint is transferred to the other party for automatic verification using a specialized transcriber based on a dictionary of phonetically distinct words. The session is accepted if the fingerprints match (and the speaker's voice is validated by the user).

any case, the fingerprint generation, transfer, and comparison happens transparently, without user involvement.

Figure 3 shows a high-level overview of our approach. Stethoscope does not require the users to speak the fingerprint for each new session, rather it asks the users to speak all the words in the fingerprint dictionary only once to build the fingerprint atomic unit dictionary. Stethoscope then uses this dictionary as an input to a limited domain text-to-speech tool to produce valid fingerprint in the user's voice for each session. Similar to the work of [31], Stethoscope automatically compares the fingerprints. However, unlike [31], to improve the transcription accuracy the fingerprint dictionary is picked from phonetically distinct words for even and odd positions.

Relieving the user from speaking and comparing the fingerprint has several important benefits:

Avoiding opt-out behavior: Since the fingerprint can be produced and transferred to the other party without the user's involvement, the fingerprint transfer can happen transparently even if one party is not willing to run the fingerprint validation protocol. In fact, part of the security improvement in our approach comes from the fingerprint exchange automation which directly impacts the usability. By making the process more transparent to the users and by reducing the burden on them (in speaking and comparing fingerprints), we hope to prevent them from opting out from these security protocols. Note that if the fingerprint validation is not to be performed (as users may not be willing to speak and share the fingerprints), all security offered by Crypto Phones is lost. Therefore, we target the security issues in Crypto Phones from the perspective of enforcing and automating the fingerprint comparison.

Reducing user errors and skip-through: Possible errors in reading the fingerprint and/or speaking the fingerprint would be reduced. Such errors in speaking the words may result in rejection or prolongation of a benign session establishment, which requires the users to re-execute the fingerprint validation protocol and may impact the usability of the system. Moreover, by automating the fingerprint comparison process, we reduce the chances of users skipping-through or rushing through the process without paying much attention (e.g., hitting "accept" without really comparing the fingerprint), which in turn may lead to acceptance of attacked sessions.

Reducing user burden and number of user tasks: Current Crypto Phone deployments require the users to "speak the fingerprint", "verify the correctness of fingerprint", and "verify the speaker", for each session. By automatically generating the audio fingerprint, transparently transferring the fingerprint, and automatically comparing the fingerprint the user would be responsible only for verifying the speaker, which could help to more accurately verify the speaker. Therefore, automation due to speech synthesis and transcription could potentially help to improve the security by reducing errors in verifying manual/automatic impersonated voices. Our approach could therefore significantly improve the security and the usability of Crypto Phones by automating fingerprint comparison to fully eliminate the user's role in the fingerprint validation protocol, leaving the only optional task of speaker verification for the user. This single tasking, can also improve the speaker verification since the cognitive load on the users is reduced. Moreover, such automation allows the use of longer fingerprint since users are no more obligated to speak and to compare the fingerprints, thereby, the security of the fingerprint validation protocol against MITM attacks can be improved.

It is important to note that in Stethoscope, the user is not completely taken out of the loop of fingerprint validation — the receiving side user still verifies the voice of the sending side user. Simply sending the fingerprints in any audio encodings makes the system vulnerable to sophisticated (yet possible) voice MITM attacks. Voice authentication is a crucial element to establish source authentication in Crypto Phones in the desirable absence of centralized mechanisms such as TLS.

D. System Assumptions and Threat Model

Our system assumes the same threat model as in traditional Crypto Phones. In this model, the devices running the protocol are assumed to be trusted. Since the threat model of Crypto Phones assumes that the communicating devices themselves are secured and uncompromised, storing the audio dataset on the device does not impose any risk. That is, we assume that the attacker cannot steal the fingerprint dataset from the users' devices and therefore cannot create fingerprint audio samples in the user's voice. Essentially, if the attacker accesses the devices the encryption keys would be leaked and there would be no more need for stealing the fingerprint dictionary. Therefore, storing the dataset on the devices offers the same level of security as storing the keys using which the devices secure their communication.

Although there is no evidence of running a voice MITM on Crypto Phones in the wild, in case such attacks happen in future, we offer a higher level of security as Crypto Phones, and a similar level of security as semi-automated fingerprint comparison [31] against the voice MITM attack and we do not intend or claim to address such issues. However, since Stethoscope reduces the user tasks, perhaps the performance of the users in recognizing attacked sessions would improve.

Finally, since we assume that the devices running the protocol are to be trusted, we also assume that the incorporated transcribers are not malicious or under influence of the attacker's adversarial training samples. Therefore, the automated speech-to-text tool is assumed to reliably transcribe the audio samples. This assumption is also similar to the one in [31].

IV. Stethoscope Evaluation Study Design

A. Objectives

To evaluate 1) the accuracy of the automated speech-to-text tool in transcribing fingerprint derived from PGP word list, and 2) the accuracy of the users in manually verifying a speaker's voice and detecting a converted voice, we designed a study with the following goals.

1) **How well does Stethoscope perform at automatically comparing the fingerprint?** We first report on the transcriber word error rate (WER), representing the difference between a given reference fingerprint and the transcribed fingerprint. We also report *False Positive Rate* of fingerprint comparison (FPR_{cc}) which denotes the probability of accepting a mismatching fingerprint due to potential transcription errors. Such instances lead to acceptance of a data MITM attack. Finally, we report *False Negative Rate* of fingerprint comparison (FNR_{cc}) representing the probability of rejecting a valid fingerprint by the transcriber, which may impact the usability of the system.
2) **How often do the users accept a fingerprint spoken in a speaker's voice or in a synthesized voice?** For security assessment against the voice MITM attack, we report on *False Positive Rate* of Speaker Verification (FPR_{sv}) representing the probability of accepting a machine synthesized voice by the users. Also, to evaluate the accuracy of the users in detecting the Stethoscope automatically generated fingerprint, we report on *False Negative Rate* of Speaker Verification (FNR_{sv}) representing the probability of the user rejecting a valid speaker.
3) **How does fingerprint size affect the error rates?** We are interested in quantifying the changes in the error rates with the increase in the fingerprint size. In our study, we measure this effect for 4-word and 8-word fingerprint.
4) **In comparison with traditional Crypto Phones, what improvements does Stethoscope provide?** As a baseline for our study, we intend to compare the performance and accuracy of Stethoscope with traditional Crypto Phones.

B. Study Setup

PGP Word List: As the fingerprint dictionary we use PGP word list, which is a list of 512 phonetically distinct words, distributed in two sets of 256 words each. This list is used by several security applications to map data bytes to words whenever the users communicate a byte array over the voice channel. Each byte is mapped to one single word that has an optimal phonetic distance with other words. Therefore, changing a single bit in the byte string results in a different word representation. Important to our application, having separate lists for even and odd positions could help to detect any mistaken deletion or insertion of the words by the transcriber. Besides, the distance of the words could potentially help to reduce the transcription errors.

Voice Dataset: We asked two male and two female speakers to record their non-noisy voice reciting the PGP words using an audio recording software on their own personal computer or device. The average time to recite the PGP words took about 8 minutes for each speaker. As mentioned earlier, in practice the user does not need to read all the words at the same time and can build the dictionary gradually. We considered the collected samples as the original speaker of the study (victims of the attack). Based on the collected data, we created our original speaker audio dataset, consisting of 10 samples of 4-word and 10 samples of 8-word fingerprints in each of the four speakers' voice. The samples were used to evaluate the accuracy of the automated transcription and to evaluate the accuracy of the users in verifying the original speaker's voice.

To evaluate the accuracy of the users in detecting the synthesized voice generated, we used the voice conversion technique. We used audio of each of the female/male speakers as the victim (target of the voice conversion), and voice of the other female/male speaker as the attacker (source of the conversion). We trained Festvox voice transformation tool [9] to convert the attacker's voice to the victim's voice. This type of voice synthesis was used in [29]–[31] to perform the conversion attack against Crypto Phones.

During the voice conversion training phase, we used all the collected data (the PGP words spoken by the attacker and the victim). The average duration of the training audio was about 8 minutes, consisting of 40 audio files. During the testing phase, we used our original speaker audio dataset consisting of the 10 samples of 4-word and 10 samples of 8-word fingerprints (used to evaluate the transcription and original speaker's voice) as the source of the conversion (the attacker's voice) and converted them to the victim's voice. This collection of total 80 samples in the original user's voice (4 original speakers, 10 4-word, and 10 8-word) fingerprint, and 80 samples in the converted voice (4 converted voice of the attacker, 10 4-word, and 10 8-word) was used to evaluate the accuracy of manual speaker verification in verifying the original speaker and the converted voice. The audio samples were all adjusted to 16kHz, 16bit mono format for compatibility with Festvox.

VoIP System: To create realistic VoIP-quality audio samples (to evaluate the transcription), we set up a telephony system using FreeSWITCH [10] open source telephony platform. We setup FreeSWITCH 1.6.20 on a Google Cloud Platform virtual server[2]. We configured the FreeSWITCH with standard features to allow voice communication between two clients. Our VoIP client was developed in C# based on OZEKI VoIP SIP SDK [14]. Using this SDK we registered two VoIP clients to the FreeSWITCH server and streamed the fingerprint audio files as voice into the voice call established between the two clients. On the receiver's side, we recorded the call (the streamed fingerprint) and stored it in an audio file, which was later fed into the speech-to-text transcription to evaluate the performance of the transcriber.

C. Speech-to-Text Transcription

To be able to compare our results with the one presented in [31], we used IBM Speech to Text API [16] to transcribe the audio samples. IBM Speech to Text transcribes audio in real-time and can be customized for various contents for improved accuracy. In our application, this could help to customize the service specifically for PGP word list.

We created an instance of the service on IBM Cloud platform and created the credential to authenticate to this instance.

[2]The server runs Debian 8 (Jessie) on a g1-small (1 vCPU, 1.7 GB memory) machine with Intel Ivy Bridge CPU platform in us-central1-f zone.

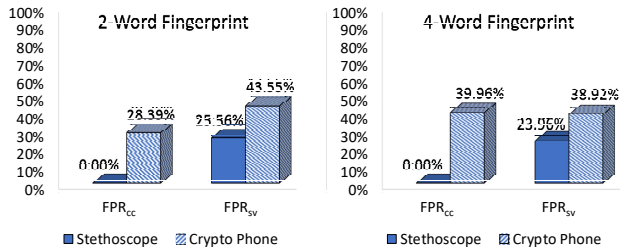

Fig. 4: A comparison between the FPR of Stethoscope and the traditional Crypto Phone (as reported in [30])

We developed a Java application that connects to the service and sends cURL commands to transcribe the audio samples with three alternatives. We parsed the received JSON response and compared the transcript with the manually transcribed corpus (the reference corpus) to report on the transcriber's WER, FPR$_{cc}$, and FNR$_{cc}$.

D. Speaker Verification Protocol Flow

To assess the accuracy of the users in verifying original speaker and converted voice, we designed a study website using LimeService survey platform and recruited 36 participants through Amazon MTurk crowd sourcing platform. This study was approved by our IRB and the participation in the study was strictly voluntary. The participants were compensated $2 for their effort. We informed the participants about the goal of the study and provided instruction on how to proceed with the survey. We informed the participants about Cyrpto Phone fingerprint validation functionality, and the importance of speaker verification in defeating MITM attacks.

We asked the participants in the study to fill out a demographic information questionnaire. For each of the speakers in the study, we first played a 1-minute voice sample of the speaker reading words in the PGP list and asked the participants to get themselves familiar with the voice. Followed by the familiarization with each voice, we randomly presented 40 questions related to the 4-word and 8-word, original and converted voice, for that particular speaker. Each of the 40 questions presented an audio sample and asked the participant to whether they found the audio sample to be of the same speaker they got familiar (answer options being Yes/No/Uncertain). After answering the 40 questions for one speaker, the survey webpage guided the participant to the familiarization and questions related to the next speaker.

V. Analysis and Results

A. Results of Automated Fingerprint Comparison

We set up an IBM speech-to-text service and developed a program to query the service for transcription of the 80 audio files, consisting of 480 words (10 samples of 4-word fingerprints and 10 samples of 8-word fingerprints, spoken by 2 male and 2 female speakers).

We first report the word error rates of the IBM speech-to-text service. In transcribing the 480 words in our dataset, the transcriber made 65 errors, that is the WER for transcribing our dataset consisting of PGP word list is estimated to be around 13.65%. To report the FPR$_{cc}$, implying the success of the data MITM attack, we manually reviewed all the instances of incorrectly transcribed words and noticed that none of the errors lead to the generation of a valid fingerprint. This shows that if the fingerprint dictionary includes only phonetically distinct words (as is the case for the words in the PGP list), the FPR$_{cc}$ is **0%**, i.e., the system would not accept any possible attacked fingerprint. A similar achievement was reported in [31] for the semi-automated fingerprint comparison model, even though Shirvanian et al. [31] did not deploy a phonetically distinct dictionary of words. In fact, the motivation behind using automated transcription is to eliminate the human errors in comparing the fingerprint and thereby reducing the chance of accepting an attacked session, which is reiterated to be successful in our study. Note that FPR$_{cc}$ for manual comparison is reported to be above 30% [30]. Therefore, similar to the work of [31], Stethoscope significantly improves the security.

Since the WER of the transcriber is about 13%, in several of the valid fingerprint samples one or more words were transcribed incorrectly. Although the transcription errors do not replace any of the words with other words in the same dictionary, the errors lead to rejection of valid fingerprint samples. Similar to the study evaluation approach of [31], we use the FNR$_r$ metric (r stands for relaxed) to evaluate the accuracy of the transcriber. This definition of relaxed in FNR$_r$, allows acceptance of a transcribed audio even if half of the words in the fingerprint are transcribed incorrectly and none of the incorrectly transcribed words belongs to the fingerprint dictionary. With this definition, the FNR$_r$ for a Stethoscope generated fingerprint is **0%**, meaning all the fingerprint samples in our dataset are transcribed such that at least half of the words are transcribed correctly. FNR$_r$ for [31] was reported to be 4.38% for 4-word fingerprints and 7.44% for 8-word fingerprints due to the choice of the dictionary.

The relaxed mode changes the theoretical security of the underlying fingerprint protocol to $2^{-k/2}$ for a k-bit fingerprint (from 2^{-k}). Therefore, we have to double the size of the fingerprint to achieve the same level of theoretical security as traditional designs. For example, in comparing the results of Stethoscope with traditional Crypto Phones, FNR of 2-/4-word fingerprints in traditional Crypto Phone should be compared with 4-/8-word fingerprints in Stethoscope. For traditional Crypto Phones, the FNR was reported to be 22.31% and 25.00% for 2-/4-word fingerprints, respectively. As the result suggests, Stethoscope can decrease the error rate over both traditional and the semi-automated fingerprint comparing designs of Crypto Phones by eliminating FNR$_r$, due to the use of phonetically distinct words in the dictionary. This relaxation does not degrade the security eventually, but in fact, we can achieve a 0% FNR and 0% FPR, which is a significant practical security improvement over traditional designs. The differences between the error rates in Stethoscope with a 2/4-word fingerprint and the error rates in traditional Crypto Phone with a 4/8-word fingerprint are illustrated in Figure 4.

B. Results of Manual Speaker Verification

We recruited 36 participants through Amazon MTurk platform to evaluate the performance of the users in verifying Stethoscope generated fingerprint samples, and synthesized fingerprint samples. We present the results of the study next.

TABLE II: Results of the Speaker Verification for Original Speaker fingerprint samples generated by Stethoscope

	No Hearing Impairment			Hearing Impairments			All the Participants		
Speaker 1 (Female)		**4-word**	**8-word**		**4-word**	**8-word**		**4-word**	**8-word**
	Yes	62.96%	57.41%	**Yes**	57.78%	44.44%	**Yes**	61.67%	54.17%
	No	23.33%	25.93%	**No**	15.56%	18.89%	**No**	21.39%	24.17%
	Uncertain	13.70%	16.67%	**Uncertain**	26.67%	36.67%	**Uncertain**	16.94%	21.67%
Speaker 2 (Male)		4-word	8-word		**4-word**	**8-word**		**4-word**	**8-word**
	Yes	75.56%	79.63%	**Yes**	54.44%	53.33%	**Yes**	70.28%	73.06%
	No	12.22%	10.37%	**No**	14.44%	11.11%	**No**	12.78%	10.56%
	Uncertain	12.22%	10.00%	**Uncertain**	31.11%	35.56%	**Uncertain**	16.94%	16.39%
Speaker 3 (Female)		**4-word**	**8-word**		**4-word**	**8-word**		**4-word**	**8-word**
	Yes	65.56%	67.78%	**Yes**	56.67%	55.56%	**Yes**	63.33%	64.72%
	No	21.11%	18.15%	**No**	13.33%	15.56%	**No**	19.17%	17.50%
	Uncertain	13.33%	14.07%	**Uncertain**	30.00%	28.89%	**Uncertain**	17.50%	17.78%
Speaker 4 (Male)		**4-word**	**8-word**		**4-word**	**8-word**		**4-word**	**8-word**
	Yes	81.11%	76.67%	**Yes**	62.86%	72.86%	**Yes**	75.00%	74.72%
	No	8.89%	11.48%	**No**	10.00%	10.00%	**No**	10.00%	11.39%
	Uncertain	10.00%	11.85%	**Uncertain**	27.14%	17.14%	**Uncertain**	15.00%	13.89%
Overall		4-word	8-word		**4-word**	**8-word**		**4-word**	**8-word**
	Yes	71.30%	70.37%	**Yes**	57.94%	56.55%	**Yes**	67.57%	66.67%
	No	16.39%	16.48%	**No**	13.33%	13.89%	**No**	15.83%	15.90%
	Uncertain	12.31%	13.15%	**Uncertain**	28.73%	29.56%	**Uncertain**	16.60%	17.43%

1) Demographic Information: There were 69.44% males and 30.56% females among the 36 participants in our study. Most of the participants were between 18 and 44 years old (19.44% 18-24 years, 58.33% 25-34 years, 13.89% 35-44 years, and 8.33% 45-and older). 5.56% of participants were high school graduates, 13.89% had a college degree, 11.11% had Associate degree, 58.33% had Bachelor's degree and 11.11% had Master's degree. Participants were from different industries, including technology, finance, transportation, manufacturing, education, and healthcare. 38.89% of participants declared their general computer skill as Excellent, 50.00% as Good and 11.11% as Fair. With respect to general computer security background, 22.22% declared their knowledge as Excellent, 58.33% as Good, and 19.44% as Fair. This analysis shows that participants represent a diverse population by gender, age, and education. Relevant to the speaker verification task, we asked participants about any hearing impairment they might have. 75% declared that they did not have any hearing impairment, while 25% had some form of impairment.

2) Results of Speaker Verification – Original Speaker: To recall, we presented the participants with a total number of 80 audio samples, consisting of 10 samples of 4-word fingerprints and 10 samples of 8-word fingerprints spoken by two female and two male speakers. Table II shows the results of the original speaker verification for the four speakers. Since 25% of the participants had some form of hearing impairment, we divide the participants into two groups, those with hearing impairment, and those without hearing impairment and report the results for both groups, as well as the overall results averaged over all the participants and the four speakers.

The results show that after getting familiar with the voice of the speaker, those with no hearing impairment could recognize the samples generated by the Stethoscope fingerprint generation tool with about 63% to 81% accuracy. We did not notice any statistically significant differences between the results for the 4- and 8-word samples[3]. Participants with the hearing impairment had more errors in recognizing the original speakers. The results averaged over the 4 speaker shows about 70% accuracy of the participants with no hearing impairment and around 57% for those with hearing impairment.

If we remove the "Uncertain" answers, the FNR_{sv} when averaged over all the participants and the 4 speakers is 15.83% for 4-word and 15.90% for the 8-word fingerprints. The reported results in [31] shows an FNR_{sv} of 25.76% and 21.72% for 4-word and 8-word fingerprint, respectively. The FNR_{sv} for a 2-word fingerprint in the traditional Crypto Phone design with a similar study setup as our study was reported 48.33% [29]. This shows that Stethoscope can improve the speaker verification compared to the traditional Crypto Phones and and the semi-automated fingerprint comparing design of [31], perhaps due to single tasking and use of longer fingerprints.

3) Results of Speaker Verification – Converted Voice: To recall,, we presented the participants with 80 audio samples, consisting of 10 samples of 4-word fingerprints and 10 samples of 8-word fingerprints synthesized by converting the attacker's voice to the victim's voice. For each of the female/male speakers as the victim (target of the conversion) we used the voice of the other female/male speaker as the attacker (source of the conversion). After familiarization we challenged the users to recognize converted samples (the challenges were ordered randomly along with the original samples).

Table III shows the results of speaker verification for the converted voice verification for the four speakers. Similar to the original speaker, we divide the participants into two groups, with and without hearing impairment, and report the results for both groups, as well as the overall results averaged over all the participants and the four speakers.

The results show that the participants with no hearing impairment could recognize the converted voice samples with about 32% to 68% accuracy. We did not notice any statistically significant differences between the results for the 4-word and 8-word samples. The accuracy of the participants with hearing impairment is about 33% to 59% (almost similar to the participants with no hearing impairments). The results averaged over the 4 speakers show about 50% accuracy of the participants with no hearing impairment and around 46% for the participants with hearing impairment. This result is significantly different from the one obtained from verifying original speaker's voice. A comparison between the two studies

[3]All results of statistical significance of our data analysis are reported at a 95% confidence level. The Wilcoxon Singed-Rank Test is used to examine if any statistical difference has occurred.

TABLE III: Results of the Speaker Verification for Converted Voices

	No Hearing Impairment			Hearing Impairments			All the Participants		
Speaker 1 (Female)		**4-word**	**8-word**		**4-word**	**8-word**		**4-word**	**8-word**
	Yes	21.85%	23.33%	**Yes**	41.11%	27.78%	**Yes**	26.67%	24.44%
	No	41.85%	45.19%	**No**	33.33%	43.33%	**No**	39.72%	44.72%
	Uncertain	36.30%	31.48%	**Uncertain**	25.56%	28.89%	**Uncertain**	33.61%	30.83%
Speaker 2 (Male)		**4-word**	**8-word**		**4-word**	**8-word**		**4-word**	**8-word**
	Yes	18.15%	16.30%	**Yes**	21.11%	21.11%	**Yes**	18.89%	17.50%
	No	55.93%	58.89%	**No**	52.22%	52.22%	**No**	55.00%	57.22%
	Uncertain	25.93%	24.81%	**Uncertain**	26.67%	26.67%	**Uncertain**	26.11%	25.28%
Speaker 3 (Female)		**4-word**	**8-word**		**4-word**	**8-word**		**4-word**	**8-word**
	Yes	44.44%	40.74%	**Yes**	30.00%	21.11%	**Yes**	40.83%	35.83%
	No	31.48%	35.19%	**No**	35.56%	38.89%	**No**	32.50%	36.11%
	Uncertain	24.07%	24.07%	**Uncertain**	34.44%	40.00%	**Uncertain**	26.67%	28.06%
Speaker 4 (Male)		**4-word**	**8-word**		**4-word**	**8-word**		**4-word**	**8-word**
	Yes	12.59%	14.81%	**Yes**	21.43%	22.86%	**Yes**	15.83%	18.06%
	No	68.89%	64.81%	**No**	58.57%	54.29%	**No**	63.61%	60.00%
	Uncertain	18.52%	20.37%	**Uncertain**	20.00%	22.86%	**Uncertain**	20.56%	21.94%
Overall		**4-word**	**8-word**		**4-word**	**8-word**		**4-word**	**8-word**
	Yes	24.26%	23.80%	**Yes**	28.41%	23.21%	**Yes**	25.56%	23.96%
	No	49.54%	51.02%	**No**	44.92%	47.18%	**No**	47.71%	49.51%
	Uncertain	26.20%	25.19%	**Uncertain**	26.67%	29.60%	**Uncertain**	26.74%	26.53%

shows that the participants are generally more successful in detecting an original speaker's voice compared to a converted voice. Besides, as can be noticed, participants seem more confident recognizing an original speaker's voice compared to the converted voice, i.e., the fraction of overall "Uncertain" answers has increased for the converted voice samples compared to the original voice samples. Overall, the results suggest that the participants are able to distinguish between the original speaker's voice and the converted voice.

If we remove the "Uncertain" answers the FPR_{sv} when averaged over all the participants and the 4 speakers is 25.56% 23.96%, for 4-word and 8-word fingerprints, respectively. The FPR_{sv} was reported to be 43.55% for a 4-word and 38.92% for an 8-word fingerprint in the traditional Crypto Phone design [30]. This shows that Stethoscope can improve the robustness of speaker verification against voice MITM attacks compared to the traditional Crypto Phone design, again perhaps due to single tasking and use of longer fingerprints. The FPR_{sv} acheived in our approach is slightly different from the one reported in [31] (18.43% and 20.45% for 4-word and 8-word fingerprints, respectively). This difference seems to stem from the quality of the recordings rather than user's intrinsic capability to detect conversion voices. While in the previous semi-automated fingerprint comparing design of [31], CMU_ARCTIC [24] professionally recorded audio dataset was used as the source and target of the voice conversion system, we used audio samples recorded by personal devices as the voice conversion training and testing datasets, which may impact the converted sample quality. Besides, in [31], authors have only evaluated samples of one male speaker, while here we evaluated the performance of speaker verification for multiple speakers. Comparing the results for various speakers also shows that accuracy is highly related to the speaker's samples. For example, the accuracy for the samples of the two male speakers seems to be higher than the one for the two female speakers. In fact for some of our speakers, the results seems to outperform the one reported in [31].

C. Study Limitations

We examined the performance of a cloud-based speech-to-text tool in transcribing the PGP word list with a set of 4 speakers (a total number of 2048 words). The goal of this test is not to evaluate any specific transcriber, but to show that, first, the speech-to-text tools can successfully transcribe the fingerprint generated by the text-to-speech tool, and second, to demonstrate that if a list of sufficiently distinct words is to be used, error rates in verifying the fingerprint (FNR_{cc} and FPR_{cc}) is significantly lower than the error rates of manual fingerprint comparison. A larger set of audio samples, different audio formatting and sampling rate, and a diverse population of speakers with different accents could show the accuracy of the transcribers better. However, studying the performance of speech-to-text tools is out of the scope of this paper. Also, we do not intend to recommend any specific cloud-based or phone-based transcriber. In fact, any reliable transcriber can be incorporated as mentioned in our threat model. We also did not intend to study the usability of the proposed system. Since the only human task in Stethoscope , is the speaker verification, we believe there is no requirement for such evaluation, as Stethoscope does not add any extra burden on the users compared to the previous models. Hence,the system usability may be inferred from the usability of the traditional Crypto Phones and CCCP design, bearing in mind that Stethoscope, in fact, automates the majority of human tasks.

VI. Discussion and Future Directions

Integrating with Real-World Systems. Our study showed that Stethoscope is a practical and feasible approach that can effectively avoid the skip-through issue and eliminate the data MITM attacks using automated fingerprint comparison with enhanced fingerprint word dictionary. In fact, if data MITM is the only viable attack against Crypto Phones (as is currently the case in practice) integration of Stethoscope in the existing systems (e.g., WhatsApp, Viber, Silent Circle) can provide a highly secure design by automating the fingerprint transfer and comparison. Since the user is not required to verify the speaker assuming that the voice MITM attack is not practical (like in many real-world Crypto Phone apps), Stethoscope can automatically establish a call between two users, generate and transfer their respective fingerprint silently, and automatically compare the fingerprints for the purpose of data MITM attack detection. All these fingerprint validation operations can happen offline with no involvement from the user. In our future work, we plan to integrate Stethoscope with current Crypto Phone designs.

Other Transcription-based Fingerprint Validation Models. It may be possible to further improve the security and usability of Stethoscope by combining it with Automated Speaker Verification (ASV) or Voice Biometrics. In our future work, we plan to study this model, which replaces the human with the machine in the Speaker Verification task to complete the automation circle of the fingerprint validation process (automated fingerprint generation, transfer, and verification). The proposed system has the benefit of being fully automated, which would perhaps improve the user experience. In this setting, the user would not have any role and responsibility in establishing the secure connection. The user would just make phone calls, and the system runs the fingerprint validation protocol on behalf of the user. Moreover, unlike the manual schemes, it will not suffer from the problem of click-through behavior. A careful future investigation is necessary to study ASV models incorporated into Stethoscope system.

VII. Conclusions

In this paper, we propose Stethoscope, a Crypto Phone fingerprint validation model to ease the fingerprint validation process for the users and thereby, improve the security of the system against data MITM attacks. In this approach, the words in the fingerprint dictionary are spoken only once by the users. The system then uses automated fingerprint generation technique based on a limited domain text-to-speech technology to automatically generate and transfer the fingerprint over the voice call channel. Further, Stethoscope offers a highly robust automated fingerprint transcription method based on a dynamic and phonetically distinct dictionary to facilitate the fingerprint comparison process and enhance the security of the system. We built a fingerprint audio dataset and ran a study, to evaluate the effect of Stethoscope on automated fingerprint comparison and manual speaker verification. Our results showed that Stethoscope can offer a 0% FPR and 0% FNR of automated fingerprint comparison while improving the speaker verification over current Crypto Phone deployment.

Ackowledgments

We thank anonymous PST'19 reviewers for their constructive comments and guidance. We are also thankful to Dr. Ali Gezer and all members of the UAB SPIES lab for feedback on previous versions of this paper.

References

[1] "How The NSA Pulls Off Man-In-The-Middle Attacks: With Help From The Telcos," 2013, https://goo.gl/Kg4ysn.

[2] "Infosecurity - Microsoft Expands Encryption to Foil Government Snooping," 2013, http://goo.gl/Ta4H0x.

[3] "NSA and All Major Intelligence Agencies Can Listen in to Encrypted Cell Phone Calls," 2013, http://goo.gl/KJgoIv.

[4] "T-Mobile Wi-Fi Calling App vulnerable to Man-in-the-Middle attack," 2013, http://goo.gl/EQ0gT3.

[5] "Festival," 2014, http://www.cstr.ed.ac.uk/projects/festival/.

[6] "iOS and Android OS Targeted by Man-in-the-Middle Attacks," 2014, https://goo.gl/R3KW40.

[7] "ZORG - An Implementation of the ZRTP Protocol," 2015, http://www.zrtp.org/.

[8] "Google's speech recognition technology now has a 4.9% word error rate," http://bit.ly/2J9ypvg, 2017.

[9] "TRANSFORM: Flexible Voice Synthesis Through Articulatory Voice Transformation," 2017, http://festvox.org.

[10] "FreeSWITCH," https://freeswitch.org, 2018.

[11] "Google Cloud Text-to-Speech," http://bit.ly/2GZCLlg, 2018.

[12] "Lyrebird - Create a digital copy of voice," https://lyrebird.ai/, 2018.

[13] "Microsoft Speech Synthesis API," goo.gl/PfYs6H, 2018.

[14] "OZEKI VoIP SIP SDK – High performance VoIP SDK for .Net Developers," http://www.voip-sip-sdk.com/, 2018.

[15] "Silent Circle," 2018, https://silentcircle.com/.

[16] "Speech to Text — IBM Watson Developer Cloud," 2018, www.ibm.com/watson/developercloud/speech-to-text.html.

[17] "Viber Encryption Overview," 2018, https://www.viber.com/en/security-overview.

[18] "WhatsApp Security," 2018, https://www.whatsapp.com/security/.

[19] D. Akhawe and A. P. Felt, "Alice in warningland: A large-scale field study of browser security warning effectiveness," in *Proceedings of the 22th USENIX Security Symposium*, 2013.

[20] A. W. Black and K. A. Lenzo, "Limited Domain Synthesis," Tech. Rep., 2000.

[21] M. Cagalj, S. Capkun, and J.-P. Hubaux, "Key agreement in peer-to-peer wireless networks," *Proceedings of the IEEE*, vol. 94, no. 2, pp. 467–478, 2006.

[22] N. Carlini, P. Mishra, T. Vaidya, Y. Zhang, M. Sherr, C. Shields, D. Wagner, and W. Zhou, "Hidden voice commands," in *25th USENIX Security Symposium (USENIX Security 16), Austin, TX*, 2016.

[23] S. Jarecki and N. Saxena, "Auhenticated Key Agreement with Key Re-Use in the Short Authenticated Strings," in *Conference on Security and Cryptography for Networks (SCN)*, September 2010.

[24] A. W. B. John Kominek, "CMU ARCTIC Databases for Speech Synthesis," 2003. [Online]. Available: http://festvox.org/cmu_arctic/cmu_arctic_report.pdf

[25] C. Kuo, J. Walker, and A. Perrig, "Low-cost manufacturing, usability, and security: an analysis of bluetooth simple pairing and wi-fi protected setup," in *International Conference on Financial Cryptography and Data Security*, 2007.

[26] S. Laur and K. Nyberg, "Efficient mutual data authentication using manually authenticated strings," in *Cryptology and Network Security (CANS)*, 2006.

[27] S. Pasini and S. Vaudenay, "An Optimal Non-Interactive Message Authentication Protocol." in *CT-RSA*, 2006.

[28] N. Roy, S. Shen, H. Hassanieh, and R. R. Choudhury, "Inaudible voice commands: The long-range attack and defense," in *15th USENIX Symposium on Networked Systems Design and Implementation*, 2018, pp. 547–560.

[29] M. Shirvanian and N. Saxena, "Wiretapping via mimicry: Short voice imitation man-in-the-middle attacks on crypto phones," in *ACM CCS 2014*, 2014.

[30] ——, "On the security and usability of crypto phones," in *Proceedings of the 31st Annual Computer Security Applications Conference*. ACM, 2015.

[31] ——, "CCCP: Closed Caption Crypto Phones to Resist MITM Attacks, Human Errors and Click-Through," in *Proceedings of the 2017 ACM SIGSAC CCS*, 2017.

[32] M. Shirvanian, N. Saxena, and J. J. George, "On the pitfalls of end-to-end encrypted communications: A study of remote key-fingerprint verification," in *Proceedings of the 33rd ACSAC*, 2017.

[33] E. Uzun, K. Karvonen, and N. Asokan, "Usability analysis of secure pairing methods," in *Financial Cryptography & Data Security*, 2007.

[34] S. Vaudenay, "Secure communications over insecure channels based on short authenticated strings," in *Advances in cryptology–CRYPTO 2005*. Springer, 2005.

[35] G. Zhang, C. Yan, X. Ji, T. Zhang, T. Zhang, and W. Xu, "Dolphinattack: Inaudible voice commands," in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, 2017.

[36] R. Zhang, X. Wang, R. Farley, X. Yang, and X. Jiang, "On The Feasibility of Launching the Man-in-the-Middle Attacks on VoIP from Remote Attackers," in *ASIACCS*, 2009.